Autocorrelation and Relational Learning: Challenges and Opportunities

Jennifer Nevillle JNEVILLE@CS.UMASS.EDU
Özgür Şimşek OZGUR@CS.UMASS.EDU
David Jensen JENSEN@CS.UMASS.EDU
University of Massachusetts, Amherst MA 01003-9264 USA

Abstract

Autocorrelation, a common characteristic of many datasets, refers to correlation between values of the same variable on related objects. It violates the critical assumption of instance independence that underlies most conventional models. In this paper, we provide an overview of research on autocorrelation in a number of fields with an emphasis on implications for relational learning, and outline a number of challenges and opportunities for model learning and inference.

1. Introduction

Autocorrelation refers to correlation between values of the same variable on related objects. More formally, it is defined with respect to a set of related instance pairs $(z_i, z_j) \in Z$ and a variable X defined on these instances, and is the correlation between the values of X on these instance pairs. Autocorrelation is a common characteristic of many datasets. For example, hyperlinked web pages are more likely to share the same topic than randomly selected pages (Taskar et al., 2002), and proteins located in the same place in a cell (e.g., mitochondria or cell wall) are more likely to share the same function (e.g., transcription or cell growth) than randomly selected proteins (Neville & Jensen, 2002).

The prevalence of autocorrelation is not unexpected—a number of widely occurring phenomena give rise to such dependencies. Temporal and spatial locality very often result in autocorrelated observations, due to temporal or spatial dependence of measurement errors, or the existence of a variable whose influence is correlated among instances that are located closely in time or space (Mirer, 1983; Anselin, 1998). Social phenomena such as social influence (Marsden & Friedkin, 1993), diffusion processes (Doreian, 1990), and the principle of homophily (McPherson et al., 2001) give rise to autocorrelated observations as well, through their influence on social interactions that govern the data generation process.

Presence of autocorrelation is a strong motivation for relational learning and inference. It is well known that in relational domains, joint inference over an entire dataset results in more accurate predictions than conditional inference over each instance independently (Macskassy & Provost, 2003; Chakrabarti et al., 1998; Taskar et al., 2002; Yang et al., 2002; Neville & Jensen, 2003). Recent work has shown that the improvement over conditional models increases with increased autocorrelation (Jensen et al., 2004)—autocorrelation allows inferences on one object to be useful for inferences on related objects.

The presence of autocorrelation, however, also presents additional challenges for learning. A major difficulty is that the assumption of independent data instances that underlie most conventional models is no longer valid. For instance, in models constructed from temporal and spatial datasets, autocorrelation has long been recognized as a source of increased bias and variance (Anselin, 1998). These problems are only more severe in relational data that do not exhibit the regularities of temporal and spatial datasets. For example, *linkage*—a measure of the number of related instances—can be far greater and can vary dramatically throughout the dataset, and it is known that linkage interacts with autocorrelation to increase variance and such variance can bias feature selection toward features with the least amount of evidence (Jensen & Neville, 2002).

Datasets exhibiting autocorrelation are common in many fields including sociology, economics, geography, and physics (Doreian, 1990). Social network analysis often examines networks of social interactions which exhibit homophily. For example, in elementary school

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2004		**00-00-2004 to 00-00-2004**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Autocorrelation and Relational Learning: Challenges and Opportunities	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Massachusetts,Department of Computer Science,Amherst,MA,01003-9264**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Autocorrelation, a common characteristic of many datasets, refers to correlation between values of the same variable on related objects. It violates the critical assumption of instance independence that underlies most conventional models. In this paper, we provide an overview of research on autocorrelation in a number of fields with an emphasis on implications for relational learning, and outline a number of challenges and opportunities for model learning and inference.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**8**	

friendship networks, same-gender ties are more likely than different-gender ties (Anderson et al., 1999). Economic analysis often examines datasets with repeated measures of the same variable over time, which typically exhibit temporal autocorrelation (e.g., stock prices). As a consequence, researchers in these fields have investigated the effects of autocorrelation in parameter estimation, hypothesis testing, and structure search.

A common finding in these disparate fields is that departures from independence cannot be ignored—they may cause unduly complex models, and biased, inconsistent, or inefficient estimators. One possible approach is to design new statistical procedures that are robust to autocorrelation. A second one is to model dependencies explicitly.

In this paper, we provide an overview of research on autocorrelation in these fields with an emphasis on implications for machine learning. The remainder of this paper is organized as follows: First, we provide an overview of work in temporal sequence analysis focusing on work in econometrics. This field has a long history of analyzing the effects of autocorrelation. We next discuss research in spatial statistics that extend one-dimensional temporal models to address the needs of higher-dimensional spatial data, and continue with work in social network analysis on general network data. We then briefly outline models in relational learning and discuss the utility and implications of work in related fields for relational learning models.

2. Temporal Sequential Models

Linear regression models are commonly employed in both natural and social sciences to model the dependence of a single response variable Y on a set of predictor variables $\mathbf{X} = \{X^1, \ldots, X^m\}$. The conventional linear regression model is specified as follows:

$$Y_i = \beta \mathbf{X_i} + \epsilon_i \tag{1}$$

where β is a vector of weights, ϵ is a normally-distributed error term with mean 0, and i is an index over data instances. The weight vector β is usually estimated using Ordinary Least Squares (OLS), which is known to be the Best Linear Unbiased Estimator (BLUE)—the minimum variance estimator for the class of linear unbiased estimators.

One of the implicit assumptions underlying these models is that instances are independent. However, this assumption is violated in many datasets consisting of observations over time. For example, the daily closing price of a stock market index (e.g., S&P500) can be represented as a time series. It is well known that stock prices exhibit autocorrelation over time—the best prediction of tomorrow's stock prices is based on today's prices (Wooldrige, 2003). [1]

If a conventional linear regression model is used to model autocorrelated data, the residuals of the model will be autocorrelated. This violates the modeling assumption of independent and identically distributed errors. For example, if equation 1 is used to regress a number of market indicators $\mathbf{X}$ (e.g. unemployment rate, federal interest rate) on the index price Y, errors will be similar for instances close in time due to the autocorrelation of Y. Serially correlated errors can be detected using a variety of statistics. The most widely-used is the Durbin-Watson statistic, which is a normalized sum of the squared differences of successive terms in a time series (Kennedy, 1998).

When the errors are autocorrelated, OLS estimators are unbiased, but they are no longer BLUE (Wooldrige, 2003). That is, there exist other unbiased linear estimators with lower variance. Not accounting for the autocorrelation structure results in larger sampling errors for the β estimates. Typically, this increased variance will bias hypothesis tests in the direction of increased Type I errors (rejecting the null hypothesis when it is true) and will result in incorrect conclusions of significance. Furthermore, the amount of bias will increase as the level of autocorrelation increases.

Autocorrelated errors typically arise in one of two situations. First, autocorrelated errors may be due to correlated measurement errors. For example, trading patterns can produce serially correlated estimates of stock returns even when there is no serial correlation for returns in general. Returns are measured using the price of the stock on the last trade in a given time period; if the measurement time period is short and the stock is sparsely traded, the estimates of return values will exhibit autocorrelation. Models that represent such autocorrelation dependencies among error terms are known generally as *disturbances models*, but are also referred to as heterogeneity models in spatial analysis, or as serial correlation models in temporal analysis. Second, autocorrelated errors may be due to correlation of the response values. For example, as was mentioned above, the price of an index today may

[1] Unfortunately, this characteristic cannot be used for accurate prediction because the chance of a stock's future price going up is the same as it going down. The overall process is is a random walk.

influence the price tomorrow. This case is typically modeled by including a lagged value of the response variable as a regressor. Models that represent these dependencies are known generally as *effects models*, but are also referred to as autoregressive models or as dependence models in spatial and social network analysis. Below we discuss each of these in turn.

2.1. Disturbances Model

Serial correlation implies that there is systematic dependence among the error terms of individual instances. The most common form is first-order serial correlation, in which the error term in one period includes a proportion of the error term in the previous period. The is commonly referred to as an AR(1) disturbance model:

$$Y_i = \beta \mathbf{X_i} + \mu_i, \text{ where } \mu_i = \rho \mu_{i-1} + \epsilon_i \quad (2)$$

where ρ is a parameter called the autocorrelation coefficient, whose absolute value is constrained to be less than 1. When $\rho = 0$, this model reduces to the standard linear model of equation 1.

When serial autocorrelation is present, analysts generally abandon OLS in favor of Generalized Least Squares (GLS) estimators that are BLUE. Unfortunately, knowledge of the correlation structure is needed for exact GLS estimates and in general this is not known apriori. Alternative Estimated Generalized Least Squares (EGLS) methods estimate ρ and β iteratively or jointly. EGLS estimators are neither linear nor unbiased but Monte Carlo studies have shown that EGLS is preferable to OLS in many situations (Kennedy, 1998). In particular, for the AR(1) disturbances model, EGLS is equal, or superior, to OLS when $\rho > 0.3$. The most frequently used EGLS methods are Cochrane-Orcutt iterative least squares, Durbin's two-stage method, Hildreth-Lu search procedure, and maximum likelihood. These four methods mainly differ in how they estimate ρ and are asymptotically equivalent if ϵ is distributed normally. Recent studies have shown that Bayesian estimation, which averages over a number of ρ estimates, is far superior to methods that use a single estimate (Kennedy, 1998).

2.2. Effects Model

Effects models take into account dependencies among the response values by including a lagged value of the response variable as one of the regressor variables. When lag equals 1 (e.g. first-order autocorrelation), the underlying model is referred to as an AR(1) effect model:

$$Y_i = \rho Y_{i-1} + \beta \mathbf{X_i} + \epsilon_i \quad (3)$$

When the underlying process is correctly modeled with equation 3, OLS estimators are biased but consistent as long as the errors are *contemporaneously* uncorrelated. This means that the n^{th} regressor term is not correlated with the n^{th} error term; it may be correlated with other error terms. In this case, analysts consider OLS to be the most appropriate estimator (Kennedy, 1998). In small samples, the OLS estimate for ρ is downward-biased, and the OLS estimate for β is upward-biased. In general however, there are no other estimators with superior small-sample properties so analysts prefer OLS for its asymptotic properties. Research has focused on obtaining unbiased OLS estimates for a range of specific autoregressive models, with recent work proposing a Monte Carlo based approach for models with non-normal error terms, higher-order autocorrelations, and exogenous variables (Tanizaki, 2000).

If, on the other hand, the errors are contemporaneously correlated, OLS estimators are biased and inconsistent. A two-step EGLS (as described in section 2.1) is not feasible in this situation because the residuals are correlated with the exogenous variables. The most common approach to take in this situation is instrumental variable (IV) estimation, which introduces extra *instument* variables to decouple the correlation between the regressors and the error terms to produce consistent estimators.

Autoregressive conditional heteroskedasticity (ARCH) models extend the basic AR models described above to model volatility clustering with non-constant variance that depends on past information (Engle, 1995). If the model does not include lagged-dependent variables, OLS estimators are BLUE, but non-linear maximum likelihood estimators are more efficient. If the model includes lagged-dependent variables then the OLS standard errors will not be consistent. In this case, EGLS estimators are asymptotically efficient and standard errors are asymptotically valid.

3. Spatial Models

Spatial datasets are analyzed in a number of fields including geography, biology, and economics. These datasets are typically represented in discrete or continuous two-dimensional space. For example, a spatial dataset may record soil properties throughout a spatial region. Equation 1 may also be used to model these

data, for example to model the effects of soil properties on ground water contamination. In this case each vector index i indicates a point in space. We will focus on (simpler) models for discrete space where the data are represented as a lattice—each point in space corresponds to a node in the graph and is linked to a fixed number of other nodes that are closest with respect to a distance measure.

Tests for the presence of residual spatial autocorrelation are based on either OLS or ML estimates, including tests based on Moran's **I** statistic, and Wald, Likelihood Ratio, and Lagrange Multiplier tests (Anselin, 1998). If spatial data exhibit autocorrelation, the quality of OLS parameter estimates are affected in the same manner as was discussed for temporal data—OLS estimators are unbiased but they are no longer BLUE (Anselin, 1998). Again this results in biased hypothesis tests, with the amount of bias depending on the level of autocorrelation.

Dependencies among instances occur in the same manner as in the temporal model discussed above. Autocorrelated errors may be due to spatially correlated measurement errors. For example, a severe weather event may affect only part of the region, resulting in a cluster of correlated errors. On the other hand, autocorrelated errors may be due to spatial autocorrelation in the response variable itself—contamination levels are likely to correlated with the levels at nearby locations.

3.1. Disturbances Model

When the data exhibit autocorrelated disturbances, the error term of one instance influences the error terms of neighboring instances. A spatial disturbances model subsumes the first-order serial correlation model (equation 2) by allowing more general dependencies among the error terms:

$$Y_i = \beta \mathbf{X_i} + \mu_i, \text{ where } \mu_i = \rho \mathbf{W} \mu + \epsilon_i \tag{4}$$

Here $\mathbf{W}$ is an $n \times n$ weight matrix specifying the nature of dependencies among the disturbances, and ρ is the autocorrelation parameter. When $\rho = 0$ or W is uniformly 0, this model reduces to the standard linear model of equation 1. The matrix $\mathbf{W}$ is designed to represent the influence processes present in the network. Each entry w_{ij} denotes the influence node j has on node i. For example, in a first-order spatial disturbances model, row i has a value of 1 for each neighbor j of node i and all other entries are 0.

Spatial autocorrelation among error terms has been shown to affect the quality of OLS parameter estimates (Anselin, 1998). The effects are similar to those reported for temporal models—OLS estimators will be unbiased but inefficient and GLS estimators are BLUE but are of academic interest only because the correlation structure is generally unknown. Futhermore, the multidirectional nature of spatial dependencies limits the types of EGLS methods that will produce consistent estimates. Approaches based on ML or IV result in consistent estimates of ρ and therefore retain the asymptotic properties of consistency and efficiency. However, in small samples, OLS may sometimes perform equivalently, or better than EGLS, in terms of bias and mean squared error—though finite sample analysis is limited (Anselin, 1998).

3.2. Effects Model

The second type of dependency is again due to autocorrelation of the regressor values. The spatial effects model represents these dependencies with the following:

$$Y_i = \rho \mathbf{W} \mathbf{Y} + \beta \mathbf{X_i} + \epsilon_i \tag{5}$$

Again, when $\rho = 0$ or W is uniformly 0, this model reduces to the standard regression model (equation 1).

If the response variable is autocorrelated, OLS estimators will be biased, inconsistent, and inefficient regardless of the properties of the error term (Anselin, 1998). In temporal effects models (equation 3), the OLS estimates will be unbiased if the error terms show no serial correlation. The multidirectional nature of spatial dependencies however, introduces added complexity to the OLS estimates so the conditions for consistency are only met when autocorrelation is not present, when $\rho = 0$. This means that no consistent estimates can be obtained for OLS procedures, so spatial analogues of EGLS methods are not appropriate.

Maximum likelihood (ML) estimation does not suffer from the same effects that plague OLS estimation so it is the preferred method of estimation among analysts for both the disturbances and the effects model. ML estimators have attractive asymptotic properties—consistency, efficiency, normality—but are more complex and computationally intensive than OLS. We should also note here that the attractive asymptotic properties of ML estimation do not hold uniformly, but are valid under the following conditions: the existence of the log-likelihood function for the parameters, continuous differentiability of the log-likelihood func-

tion, boundedness of partial derivatives, positive definiteness and/or non-singularity of covariance matrices, and the finiteness of quadratic forms (Anselin, 1998). Typically, these conditions are satisfied if the spatial interaction structure ($\rho\mathbf{W}$) is non-explosive (i.e., the correlation between y_i and y_{i+d} goes to zero sufficiently "quickly" as $d \to \infty$, where d is graph distance).

Depending on the model form, ML estimation may involve a normalizing constant that is difficult to compute in closed form. For the models discussed above, this involves computing the log-determinant of an $n \times n$ matrix, which requires $O(n^3)$ operations for dense matrices. Research has focused on techniques to make ML estimation more tractable, including pseudolikelihood estimation (Besag, 1975), approximate ML estimation with Markov Chain Monte Carlo (MCMC) methods (Geyer & Thompson, 1992), and closed-form ML methods that avoid direct computation of the determinant (LeSage & Pace, 2001).

Hypothesis tests for ML estimates include the Wald test, the Likelihood Ratio test, and the Lagrange Multiplier test, all of which are based on the optimal asymptotic properties of the ML estimator. The tests are asymptotically equivalent but care must be taken when interpreting the tests on finite samples because some have higher Type I errors and others have higher Type II errors. The relative power of the tests for spatial data is yet to be investigated (Anselin, 1998).

4. Network Models

Spatial models have been applied extensively in the field of social network analysis where data consist of a network of interactions among entities (e.g., people, institutions). Social network datasets are represented as general graphs and differ from temporal and spatial data representations in that they are not restricted to a uniform structure. For example, to model the effects of socio-economic status on voting behavior in a community, income and status would be measured along with friendship ties to other members in the community. A set of nodes representing people and a set of edges representing their friendships forms the network graph. The graph structure varies as each person has a different number of friends. Again equation 1 may be used to model network data. In this case each vector index i indicates a node in the graph.

Spatial autocorrelation models are expressive enough to use as network autocorrelation models. Equation 4 represents a network disturbances model and equation 5 represents a network effects model (Marsden & Friedkin, 1993). In social network models, the weight matrix $\mathbf{W}$ specifies the social influence patterns present in the network and it can affect virtually all of the conclusions drawn from autocorrelation models (Leenders, 2002). Therefore, correct specification of $\mathbf{W}$ is crucial to the utility of the models. In practice, social network analysts do not estimate $\mathbf{W}$. Instead, they specify a $\mathbf{W}$ manually to model specific theories of social influence such as communication and comparison.

Social network models share the same challenges as spatial models—OLS parameter estimates of autocorrelated data will be inefficient and/or biased and inconsistent, and although ML estimates are more robust, they are computationally intensive (Doreian & K. Teuter, 1984). Simulation studies have demonstrated the superiority of ML estimates over a wide range of conditions (Doreian & K. Teuter, 1984). Although social network datasets are not restricted to a uniform structure, unfortunately there appears to be little work in social networks that examines the impact of varying graph structure on parameter estimation and hypothesis tests.

5. Models in Relational Learning

Datasets with more general dependencies than are seen in temporal, spatial, and social network data are commonplace in relational learning. For example, relational data for citation analysis can be represented as a typed, attributed graph, with nodes representing authors, papers and journals, and edges representing citation and published-in relationships. A model of paper *topic* may include attributes of related authors (e.g., speciality) and journals (e.g., prestige). However, an important characteristic of these data is that topic is autocorrelated—the topic of a paper is not independent of the topics of papers that it cites.

Relational data pose a number of additional challenges for model learning and inference. First, relational data often consider more than one type of entity in the same dataset (e.g., papers, authors and references). Second, relational data have complex dependencies, both as a result of direct relations (e.g., research paper references) and through chaining multiple relations together (e.g., papers published in the same journal). Third, relational data have varying structure (e.g., papers have different numbers of authors, references and citations).

Recent research in relational learning has produced several novel types of models to address these issues, including relational Markov network (RMNs) (Taskar et al., 2002), relational Bayesian networks (RBNs) (Friedman et al., 1999), and re-

lational dependency networks (RDNs) (Neville & Jensen, 2004). These three models have the ability to represent and reason with autocorrelation; however, only RMNs and RDNs can reason with arbitrary forms of autocorrelation—RBNs can only reason with acyclic forms of autocorrelation, such as relationships that are structured by temporal constraints (Friedman et al., 1999).

There are two major findings that relate autocorrelation to learning and inference in relational models: that autocorrelation improves joint inference, and that autocorrelation may bias feature selection. We discuss each of these below.

First, several studies have shown that *joint inference* can significantly reduce classification error (Macskassy & Provost, 2003; Chakrabarti et al., 1998; Taskar et al., 2002; Yang et al., 2002; Neville & Jensen, 2003). Joint inference refers to procedures that make simultaneous statistical judgments about the same variables for a set of related data instances. By making inferences about multiple data instances simultaneously, joint inference can exploit autocorrelation in the data—judgments about one instance can be used to improve inferences about related instances. Recent work has shown that the improvement over conditional models, which make inferences in isolation, increases with increased autocorrelation, and in general, a joint inference procedure performs better when higher-order autocorrelation is present or when few labels are known with certainty (Jensen et al., 2004). In conditional models, the utility of modeling autocorrelation depends on whether the values of the autocorrelated attributes are known. Partially labeled datasets are common, but if the known labels do not exhibit autocorrelation, they cannot be used to seed the inferences. Related work shows that the relative advantage of a joint inference procedure over a conditional procedure reduces as the percentage of labeled data increases (Macskassy & Provost, 2003).

Second, recent research has shown that autocorrelation may bias feature selection (Jensen & Neville, 2002). Concentrated linkage and autocorrelation reduce the effective sample size of a data set, thus increasing the variance of parameter estimates (e.g., feature scores) estimated using that set. This reduction in effective sample size parallels the inefficiencies in temporal and spatial estimators. As a consequence, the probability of Type I errors is increased—features formed from objects with high linkage and autocorrelation may be selected as the best feature, even when the features are random. To our knowledge, few current relational learning algorithms adjust for the increased variance in estimation. Specifically, the current instantiation of RDNs use an underlying conditional model which adjusts for this bias, but the current instantiations of RBNs and RMNs do not. Inefficient parameter estimates will impact both selective (e.g., RBN, RDN) and non-selective (e.g., RMN) models. For both types of models, the increased variance may result in overfitting. In addition, the interpretation of feature weights/scores may be more difficult for non-selective models and structure learning may be biased in selective models.

6. Summary and Discussion

Autocorrelation effects have been studied extensively in other fields and it is clear that they cannot be ignored in relational learning. In particular, if the data exhibit autocorrelation, either autocorrelated measurement errors or an autocorrelated response variable, then conventional parameter estimates will be unbiased but will have increased variance. This has implications for (1) model performance, (2) feature rankings, and (3) feature selection. When the model is learned from "small" samples, the increased variance may lead to overfitting and result in lower performance. Although, we typically have "large" datasets in relational learning, as the level of autocorrelation increases so does the variance—the amount of data needed to offset the increased variance may be be larger than we expect. Increased variance will also impact feature rankings (by feature weights/scores), and consequently feature selection. Non-selective models often use feature weights for interpretation (e.g., to identify the most important features), and selective models use feature weights to learn the structure of the model. Both these endeavors will be adversely affected by the increased variance due to autocorrelation.

How can we adjust for autocorrelation in relational models? Below, we summarize past research and discuss options for model representation, learning and inference.

6.1. Representation

The first decision is how to include autocorrelation in the model representation—whether to model autocorrelation directly through variables or indirectly in the error term. This choice corresponds to selection of the effects model, the disturbances model, or some combination of the two, and may be based on the researcher's hypothesis about the dependencies present in the data. Explicit representation may result in more interpretable models, since the influence of an autocorrelated response variable is clear. However, implicit

representation in the error term may be more broadly applicable. This approach could allow the use of existing models without a change of representation, but with only an adjustment for the effects of autocorrelation.

The second decision is how to encode the autocorrelation dependencies. This decision corresponds to the functional form of autocorrelation (e.g., first-order). For the spatial and network models discussed above, this refers to the specification of the weight matrix. For relational models, this usually refers to specification of autocorrelation features. For example, to predict the topic of a web page, we may include a feature that encodes the topics of other hyperlinked pages. While considerable attention has been paid to accurate parameter estimation in temporal and spatial autocorrelation models, it appears that researchers are less concerned with model/feature selection. However, one could imagine searching over a space of autocorrelation specifications to learn the correct structure.

6.2. Learning

The effect of autocorrelation on parameter estimation has been studied extensively. Below is a summary of the findings in temporal, spatial and network analysis:

1. If autocorrelation is ignored:
    - Parameter estimates are computationally efficient.
    - Parameter estimates are unbiased but have increased variance.
    - Hypothesis tests and confidence intervals may be biased.
2. If autocorrelation is modeled:
    - Parameter estimates are computationally complex, but more tractable approximate methods exist.
    - Parameter estimates are asymptotically optimal but may be biased in finite samples.
    - Finite sample comparison is limited. More complex estimation techniques may not always be justified.
    - Hypothesis tests are asymptotically unbiased but the relative power of various tests may vary on finite samples.

Some examples of model parameters in relational learning include clique potentials and feature weights. Results for temporal and spatial analysis indicate that there may be a tradeoff between computational efficiency and accurate parameter estimation. Understanding the effect of varying levels of autocorrelation on parameter estimation for finite samples is an important area for research for the relational learning community.

The effects on parameter estimation will also impact structure learning. Structure learning typically involves feature selection, which corresponds to either explicit or implicit hypothesis testing. It has been shown that autocorrelation can lead to increased Type I errors in hypothesis tests, which may lead to a unfair comparison among different features. The impact of these errors has not been fully explored in relational learning. Initial results indicate that they lead to overly complex models with excess structure, and may degrade model performance (Neville et al., 2003).

6.3. Inference

The literature on spatial, temporal, and social network autocorrelation models does not provide much guidance for inference because it focuses on accurate model learning rather than prediction of unobserved variables. There has, however, been preliminary work in relational learning that suggests joint inference can significantly reduce classification error, and that this reduction increases with autocorrelation. Clearly, this is an area with many open questions—e.g., Can autocorrelation be exploited to improve inference efficiency? How does autocorrelation interact with various inference procedures? How does the amount of labeled data interact with the level of autocorrelation in the dataset to determine the improvement in accuracy obtained by joint inference?

7. Conclusions

Autocorrelation is ubiquitous—datasets exhibiting autocorrelation are found in a range of fields including sociology, economics, geography, and physics—and has been studied extensively. In this paper we presented findings from econometrics, spatial statistics, and social network analysis. A common finding is that ignoring autocorrelation may result in unduly complex models, and biased, inconsistent, or inefficient estimators. The effects of autocorrelation are sometimes addressed by modeling the autocorrelation explicitly, and sometimes by using statistical procedures that are robust to these effects.

For reasons we stated earlier, we expect autocorrelation to have greater impact on relational models than on temporal, spatial, and network models. Although

the presence of autocorrelation has been widely reported for relational datasets, there has been little focus on the impact of autocorrelation on model learning and inference. The results we discuss here reveal that this is an important area for future research.

Acknowledgments

The authors acknowledge helpful comments provided by the anonymous reviewers. This research is supported under a AT&T Labs Graduate Research Fellowship and by DARPA, NSF and AFRL under contract numbers F30602-00-2-0597, EIA9983215 and F30602-01-2-0566. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation hereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements either expressed or implied, of DARPA, NSF, AFRL, or the U.S. Government.

References

Anderson, C., Wasserman, S., & Crouch, B. (1999). A p* primer: Logit models for social networks. *Social Networks*, *21*, 37–66.

Anselin, L. (1998). *Spatial econometrics: Methods and models*. The Netherlands: Kluwer Academic Publisher.

Besag, J. (1975). Statistical analysis of non-lattice data. *The Statistician*, *24:3*, 179–195.

Chakrabarti, S., Dom, B., & Indyk, P. (1998). Enhanced hypertext categorization using hyperlinks. *Proc of ACM SIGMOD98*.

Doreian, P. (1990). Network autocorrelation models: Problems and prospects. In *Spatial statistics: Past, present, and future*, chapter Monograph 12, pp. 369–389. Ann Arbor Institute of Mathematical Geography.

Doreian, P., & K. Teuter, C. W. (1984). Network autocorrelation models: Some monte carlo results. *Sociological Methods and Research*, *13:2*, 155–200.

Engle, R. (Ed.). (1995). *Arch: Selected readings*. Oxford: Oxford University Press.

Friedman, N., Getoor, L., Koller, D., & Pfeffer, A. (1999). Learning probabilistic relational models. *IJCAI* (pp. 1300–1309).

Geyer, C., & Thompson, E. (1992). Constrained monte carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society, Series B*, *54:3*, 657–699.

Jensen, D., & Neville, J. (2002). Linkage and autocorrelation cause feature selection bias in relational learning. *Proceedings of the 19th International Conference on Machine Learning*.

Jensen, D., Neville, J., & Gallagher, B. (2004). *Why collective inference improves relational classification* (Technical Report 04-27). University of Massachusetts.

Kennedy, P. (1998). *A guide to econometrics*. Cambridge, Massachusetts: The MIT Press.

Leenders, R. (2002). Modeling social influence through network autocorrelation: Constructing the weight matrix. *Social Networks*, *24*, 21–47.

LeSage, J., & Pace, R. (2001). Spatial dependence in data mining. In *Data mining for scientific and engineering applications*. Kluwer Academic Publishers.

Macskassy, S., & Provost, F. (2003). A simple relational classifier. *2nd Workshop on Multi-Relational Data Mining, KDD-2003*.

Marsden, P., & Friedkin, N. (1993). Network studies of social influence. *Sociological Methods and Research*, *22:1*, 127–151.

McPherson, M., Smith-Lovin, L., & Cook, J. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, *27*, 415–445.

Mirer, T. (1983). *Economic statistics and econometrics*. New York: Macmillan Publishing Co.

Neville, J., & Jensen, D. (2002). Supporting relational knowledge discovery: Lessons in architecture and algorithm design. *Data Mining Lessons Learned Workshop, 19th International Conference on Machine Learning*.

Neville, J., & Jensen, D. (2003). Collective classification with relational dependency networks. *2nd Workshop on Multi-Relational Data Mining, KDD-2003*.

Neville, J., & Jensen, D. (2004). *Dependency networks for relational data* (Technical Report 04-28). University of Massachusetts.

Neville, J., Jensen, D., Friedland, L., & Hay, M. (2003). Learning relational probability trees. *9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Tanizaki, H. (2000). Bias correction of olse in the regression model with lagged dependent variables. *Computational Statistics and Data Analysis*, *34*, 495–511.

Taskar, B., Abbeel, P., & Koller, D. (2002). Discriminative probabilistic models for relational data. *Proceedings of UAI-2002*.

Wooldrige, J. (2003). *Introductory econometrics: A modern approach*. South-Western College Pub.

Yang, Y., Slattery, S., & Ghani, R. (2002). A study of approaches to hypertext categorization. *Journal of Intelligent Information Systems*.